4

AD-A228 178

TESTS OF FIT BASED ON THE CORRELATION COEFFICIENT

BY

RICHARD A. LOCKHART and M. A. STEPHENS

TECHNICAL REPORT NO. 436
OCTOBER 4, 1990

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

1. INTRODUCTION.

1.1 The regression model in goodness-of-fit.

Suppose a random sample $X_1, X_2, \ldots X_n$ comes from distribution $F_o(x)$ and let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics. $F_o(x)$ may be of the form $F(w)$ with $w = (x-\alpha)/\beta$; α is then the location parameter and β is the scale parameter of $F_o(x)$. There may be other parameters in $F(w)$, for example, a shape parameter; here we assume such parameters known, but α and β are unknown. We can suppose the random sample of X-values to have been constructed from a random sample $w_1, w_2, \ldots, w_n$ from $F(w)$, by the transformation

$$X_i = \alpha + \beta w_i \ . \tag{1}$$

If the order statistics of the w-sample are $w_{(1)} < w_{(2)} < \ldots < w_{(n)}$, we have also

$$X_{(i)} = \alpha + \beta w_{(i)} \ . \tag{2}$$

Let $E(w_{(i)})$ be m_i and let v_{ij} be $E(w_{(i)} - m_i)(w_{(j)} - m_j)$; let V be the $n \times n$ matrix with entries v_{ij}. V is the covariance matrix of the order statistics $w_{(i)}$. From (2) we have

$$E(X_{(i)}) = \alpha + \beta m_i \tag{3}$$

and a plot of $X_{(i)}$ against m_i should be approximately a straight line with intercept α on the vertical axis and slope β. The values m_i are the most natural values to plot along the horizontal axis to achieve a straight line plot, but for most distributions they are difficult to calculate.

Various authors have therefore proposed alternatives T_i which are convenient functions of i ; then (2) can be replaced by the model

$$X_{(i)} = \alpha + \beta T_i + \epsilon_i \tag{4}$$

where ϵ_i is an "error" which has mean zero only for $T_i = m_i$.

A common choice for T_i is $H_i \equiv F^{-1}\{i/(n+1)\}$ or similar expressions which approximate m_i . A test of

$$H_0: \text{ the X-sample comes from } F_o(x), \tag{5}$$

can then be based on how well the data fits the line (3) or (4).

1.2 Example. As an example, suppose it is desired to test that the X-sample is normally distributed, with unknown mean μ and variance σ^2 . Then $F(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{-t^2/2}\, dt$, and the w-sample is standard normal. Then (1) becomes

$$X_i = \mu + \sigma w_i$$

and (3) is

$$E(X_{(i)}) = \mu + \sigma m_i$$

where m_i are the expected values of standard normal order statistics. For this distribution, $\alpha = \mu$ and $\beta = \sigma$.

1.3 Measures of fit.

The practice of plotting the $X_{(i)}$ against m_i (or against another set of constants T_i which approximate the m_i-values) and looking to see if a straight line results, is time-honored as a quick technique for testing normality. An improvement on this procedure by eye, is to measure how well the data fits the line (3). Three main approaches to measuring the fit can be identified. The first is simply to measure the correlation coefficient $R(X,T)$ between the paired sets X_i and T_i. A second method is to estimate the line $\alpha + \beta T_i$, using generalized least squares to take into account the covariance of the order statistics, and then to base the test of fit on the sum of squares of residuals. Finally, a third technique is to estimate β from (2) using generalized least squares, and to compare this estimate with the estimate of scale given by the sample standard deviation. In this article we explore the first two of these methods, which are often closely connected.

1.4 The correlation coefficient.

The simplest of the three methods above is to use the correlation co-efficient $R(X,T)$. Here we extend the usual meaning of correlation, and also that of variance and covariance, to apply to constants as well as random variables. Thus let X refer to the vector $X_{(1)},\ldots,X_{(n)}$, and T to vector $T_1,\ldots,T_n$; let $\bar{X} = \dfrac{\Sigma X_{(i)}}{n}$ and $\bar{T} = \dfrac{\Sigma T_i}{n}$, (all sums are for $i = 1$ to n) and define the sums

$$S(X,T) = \Sigma(X_{(i)} - \bar{X})(T_i - \bar{T}) = \Sigma X_{(i)}T_i - n\bar{X}\bar{T} ;$$

$$S(X,X) = \Sigma(X_{(i)} - \bar{X})^2 = \Sigma(X_i - \bar{X})^2 ;$$

$$S(T,T) = \Sigma(T_i - \bar{T})^2 .$$

$S(X,X)$ will often be called S^2 .

The variance of X is then $V(X,X) = \frac{1}{n-1} S(X,X)$, the variance of T is $V(T,T) = \frac{1}{n-1} S(T,T)$, and the covariance of X and T is $V(X,T) = \frac{1}{n-1} S(X,T)$. The correlation coefficient between X and T is

$$R(X,T) = \frac{V(X,T)}{\{V(X,X)V(T,T)\}^{\frac{1}{2}}} = \frac{S(X,T)}{\{S(X,X)S(T,T)\}^{\frac{1}{2}}} .$$

Statistics $R(X,m)$ (called sometimes R) or $R^2(X,m)$ are attractive statistics for testing the fit of X to the model (2), since if a "perfect" sample is given, that is, a sample whose ordered values fall exactly at their expected values, $R(X,m)$ will be 1 , and the value of $R(X,m)$ can be interpreted as a measure of how closely the sample resembles a perfect sample. Then tests based on $R(X,m)$, or equivalently on $R^2(X,m)$ will be one-tailed; rejection of H_0 occurs only for low values of R . Suppose $\hat{X}_{(i)} = \hat{\alpha} + \hat{\beta}T_i$, where $\hat{\alpha}$ and $\hat{\beta}$ are the usual regression estimators of α and β (ignoring the covariance between the $X_{(i)}$). It is possible to set up the standard ANOVA table for straight line regression:

$$\text{Regression} \quad SS = \frac{S^2(X,T)}{S(T,T)}$$

$$\text{Error} \quad SS = S^2 - \frac{S^2(X,T)}{S(T,T)} = \Sigma(X_{(i)} - \hat{X}_{(i)})^2$$

$$\text{Total} \quad SS = S^2 = S(X,X)$$

and it is clear that

$$\frac{\text{Error SS}}{\text{Total SS}} = 1 - R^2(X,T).$$

Define, for any T vector,

$$Z(X,T) = n\{1 - R^2(X,T)\}.$$

Then $Z(X,T)$ is a test statistic equivalent to $R^2(X,T)$, based on the sum of squares of the residuals after the line (3) has been fitted. $Z(X,T)$ has, in common with many other goodness-of-fit statistics e.g., chi-square, and EDF statistics, the property that the larger $Z(X,T)$ is, the worse the fit. Sarkadi [1975] and more recently Gerlach [1979] have shown consistency for correlation tests based on $R(X,m)$, or equivalently $Z(X,m)$, for a wide class of distributions including all the usual continuous distributions. This is to be expected, since for large n we expect a sample to become perfect in the sense above. We can expect the consistency property to extend to $R(X,T)$ provided T approaches m sufficiently rapidly for large samples.

1.5 Censored data. $R^2(X,T)$ can easil_ be calculated for censored data, provided the ranks of the available $X_{(i)}$ are known. These are paired with the appropriate T_i and $R^2(X,T)$ is calculated using the same formula as above, with the sums running over the known i .

For example if the data were right censored, so that only the r smallest values $X_{(i)}$ were available, the sums would run for i from 1 to r; if the data were left-censored, with the first s values missing, the i would run from $s+1$ to n . Tables of $Z(X,T)$ for $T = m$ or H , for testing for the uniform, normal, exponential, logistic, or extreme-value distributions have been published by Stephens (1986).

2. CORRELATION TESTS FOR THE UNIFORM DISTRIBUTION.

For the uniform distribution for X , between limits (a,b), written $U(a,b)$, we have $F(w) = w$, $0 < w < 1$, and $X_i = a + (b - a)W_i$; hence $\alpha = a$, $\beta = b - a$. Then $m_i = E(W_{(i)}) = i/(m+1)$; also $H_i = m_i$. The order statistics $X_{(i)}$ could be plotted against i instead of against $i/(n+1)$; the scale factor $1/(n+1)$ does not change the correlation coefficient, and $R(X,m) = R(X,H) = R(X,T)$ where $T_i = i$.

In discussing tests for the uniform distribution, we distinguish four cases:

Case 0; a,b both known;

Case 1; a unknown, but (b-a) known;

Case 2; a known, (b-a) unknown;

Case 3; both a and b unknown.

Case 0. Here a and b are both known, so that α and β are known in (1). The transformation $X' = (X-a)/(b-a)$ then reduces the problem to a test that X' is $U(0,1)$. There are of course many tests for this special case (see, eg., Stephens, 1986). In the present context, the test will be based on the residuals from the known line $F(x') = x'$, $0 < x' < 1$; that is, on the statistic $Z_0 = \Sigma\{X'_{(i)} - i/(n+1)\}^2$. It is clear that Z_0 has the same asymptotic distribution as the well-known Cramér-von Mises statistic $W^2 = \Sigma\{X'_{(i)} - (2i - 1)/(2n)\}^2 + 1/(12n)$, and, for small samples, the two statistics will have much the same power properties.

Case 1. Here the model is $X_{(i)} = a + \beta W_{(i)}$, with $\beta = b-a$ known. Substitute $X'_{(i)} = X_{(i)}/\beta$; then the model becomes $X'_{(i)} = a/\beta + W_{(i)}$, and $E(X'_{(i)}) = \alpha + (m_i - \bar{m})$, where $\alpha = a/\beta + \bar{m}$. Ordinary least squares gives $\hat{\alpha} = \bar{X}'$. Hence $\hat{X}'_{(i)} = \bar{X}' + m_i - 0.5$ and the test statistic based on residuals is $Z_1 = \Sigma\{X'_{(i)} - \bar{X}' - (m_i - 0.5)\}^2$. Z_1 has similar properties to the Watson U^2 statistic

$U^2 = \Sigma[X'_{(i)} - \bar{X} - \{(2i-1)/(2n) - 0.5\}]^2 + 1/(12n)$, and has the same asymptotic distribution.

Case 2. For Cases 2 and 3 the situation becomes much harder, and considerable analysis is required to obtain the asymptotic distributions of the test statistic $n\Sigma(X_{(i)} - \hat{X}_{(i)})^2/\Sigma(X_{(i)} - \bar{X})^2$, the denominators being necessary, and complicating the analysis, because for these two cases the scale must be estimated. We state the results and give proofs later. For Case 2, the model is $E(X_{(i)}) = a + \beta m_i$, with β unknown and a known. Set $X'_{(i)} = X_{(i)} - a$, so that $E(X'_{(i)}) = \beta m_i$, and estimate β by least squares; then $\hat{\beta} = \Sigma X'_{(i)} m_i/\Sigma m_i^2$. Thus $\hat{X}_{(i)} = a + \hat{\beta} m_i$, and the test statistic is

$Z_2 = n\Sigma\{X_{(i)} - \hat{X}_{(i)}\}^2/\Sigma\{X_{(i)} - \bar{X}\}^2$. Z_2 has the same asymptotic distribution as $Z_2^* = \Sigma v_i/\lambda_i$ where v_i, for $i = 1,2,\ldots$, are independent χ_1^2 variables; λ_i is an infinite set of positive weights given by $\lambda_i = \theta_i^2$, where θ_i are the solutions of $\tan \theta_i = \theta_i$, $\theta_i > 0$ (see Section 3 below).

Case 3. For Case 3, the model is $E(X_{(i)}) = \alpha + \beta(m_i - 0.5)$, with α,β unknown, and leas squares gives $\hat{\alpha} = \bar{X}$ and $\hat{\beta} = \Sigma[\{X_{(i)} - \bar{X}\}m_i]/\Sigma(m_i - \bar{m})^2$. The test statistic is now the correlation coefficient $R(X,m)$ or equivalently $Z_3 \equiv Z(X,m)$; $Z_3 = n\{1 - R^2(X,m)\} = n\Sigma\{X_{(i)} - \hat{X}_{(i)}\}^2/\Sigma\{X_{(i)} - \bar{X}\}^2$, where $\hat{X}_{(i)} = \hat{\alpha} + \hat{\beta}(m_i - 0.5)$. Z_3 has the same asymptotic distribution as $Z_3^* = \Sigma v_i/\lambda_i$, where, as above, v_i are independent χ_1^2 variables. The constants λ_i are positive weights given in two infinite sets:

Set 1: $\lambda_i = 4\pi^2 i^2$, $i = 1,2,\ldots$.

Set 2: $\lambda_k = 4\phi_k^2$, $k = 1,2,\ldots$, where ϕ_k are the solutions of $\tan \phi_k = \phi_k$, $\phi_k > 0$.

The derivation of the weights for Cases 2 and 3 will be given in the next section.

3. ASYMPTOTIC PROPERTIES OF $Z(X,m)$.

3.1. Case 3. It is convenient to give the asymptotic results for Case 3 (the more difficult case) first. Suppose, without loss of generality, that the sample comes from $U(0,1)$. However, the model is fitted without this knowledge: thus the fitted model is

$$X_{(i)} = \alpha + \beta(m_i - \bar{m}) + \epsilon_i \ . \tag{6}$$

As stated in Section 2, this leads to the test statistic

$$Z(X,m) = n\{1 - R^2(X,m)\} \quad = \Sigma(X_{(i)} - \hat{X}_{(i)})^2/\{\Sigma(X_i - \bar{X})^2/n\}.$$

Asymptotically, the denominator tends to 1/12; thus we must study $X_{(i)} - \hat{X}_{(i)}$. This may be written

$$X_{(i)} - \hat{\alpha} - \hat{\beta}(m_i - \bar{m}) = X_{(i)} - \bar{X} - (\hat{\beta} - 1)(m_i - \bar{m}) - (m_i - \bar{m})$$

$$= X_{(i)} - m_i - (\bar{X} - \bar{m}) - (\hat{\beta} - 1)(m_i - \bar{m}). \tag{7}$$

The terms on the right hand side of (7) can be expressed in terms of the quantile process $Q_n(t) = X_{[nt]} - m_{[nt]}$, $0 \le t \le 1$, where $[nt]$ is the greatest integer in nt. For t given by i/n, we have

$$X_{(i)} - m_i = Q_n(t);$$

$$\sqrt{n}(\bar{X} - \bar{m}) = \int_0^1 Q_n(s)ds + o_p(n^{-\frac{1}{2}})$$

$$\sqrt{n}(\hat{\beta} - 1) = \frac{n^{-\frac{1}{2}}\Sigma(X_{(i)} - \bar{X} - m_i + \bar{m})(m_i - \bar{m})}{\Sigma(m_i - \bar{m})^2/n}$$

$$= 12\int_0^1 (t - \tfrac{1}{2})\{Q_n(t) - \int_0^1 Q_n(s)ds\}dt + o_p(n^{-\frac{1}{2}}),$$

recalling that $\bar{m} = \frac{1}{2}$ and $\Sigma(m_i - \bar{m})^2/n \to 1/12$. It is convenient to define the process $Y_n(t) = Q_n(t) - \int_0^1 Q_n(s)ds$. Then insertion of the above expressions into (7) gives

$$X_{(i)} - \hat{X}_{(i)} = Y_n(t) - \int_0^1 (u - \tfrac{1}{2}) \int_0^1 12(t - \tfrac{1}{2}).Y_n(t).dt\ du + O_p(n^{-\frac{1}{2}}). \tag{8}$$

As $n \to \infty$, let $Q(t)$, $Y(t)$ be the limiting processes for $Q_n(t)$ and $Y_n(t)$ respectively. $Q(t)$ is the well-known Brownian bridge with mean $E\{Q(t)\} = 0$ and covariance $\rho_0(s,t) = \min(s,t) - st$. $Y(t)$ then has mean 0 and covariance $\rho_Y(s,t) = \min(s,t) - \frac{1}{2}s(1 - s) - \frac{1}{2}t(1-t) + 1/12$. The process $Y(t)$ **has** already been studied in connection with the Watson statistic U^2 (Watson, 1961; Stephens, 1976). For the asymptotic distribution of $Z(X,m)$ we now **need the distribution of**

$$Z^* = \int_0^1 W^2(t)\ dt, \tag{9}$$

where, from (8), we have

$$W(t) = Y(t) - \int_0^1 (u - \tfrac{1}{2}) \int_0^1 12(t - \tfrac{1}{2})\ Y(t)\ dt\ du\ . \tag{10}$$

The covariance function of $W(t)$ requires considerable algebra but the calculation is straightforward; the result may be expressed as

$$\rho_w(s,t) = \rho_0(s,t) - \psi(s)'A\psi(t) \tag{11}$$

where $\psi(s)'$ is the transpose of $\psi(s)$, and is the 2-component vector $\{(s - \frac{1}{2});\ s(1-s)(2s - 1)\}$; A is the 2×2 matrix with rows $(-\frac{1}{5}, 1)$ and $(1,0)$. The calculation of the distribution of Z^* now follows well-known lines (see, for example, Durbin , 1973, or Stephens, 1976); Z^* has the same distribution as $S = \Sigma_i\ v_i/\lambda_i$, where i runs from

1 to ∞, v_i are independent χ_1^2 variables, and where λ_i are weights, found by solving the integral equation

$$\lambda \int_0^1 f(s)\, \rho_w(s,t)\, ds = f(t) \tag{12}$$

for eigenvalues λ_i and eigenfunctions $f_i(t)$.

The solution of (12) is found as follows. The covariance $\rho_w(s,t)$ can be expressed as $\rho_w(s,t) = \min(s,t) + g(s,t)$, with

$$g(s,t) = \frac{6}{5}st - \frac{11}{10}s + 2s^2 - s^3 - \frac{11}{10}t + 2t^2 - t^3 + \frac{2}{15} - 3st^2 + 2st^3 - 3s^2t + 2s^3t.$$

Differentiation of (12) twice with respect to t then gives

$$-f(t) + 4\int_0^1 f(s)ds - 6t\int_0^1 f(s)ds - 6\int_0^1 sf(s)ds + 12t\int_0^1 sf(s)ds = \frac{1}{\lambda} f''(t). \tag{13}$$

Differentiation again gives

$$-f'(t) - 6\int_0^1 f(s)ds + 12\int_0^1 sf(s)ds = \frac{1}{\lambda} f'''(t) \tag{14}$$

and finally

$$-f''(t) = \frac{1}{\lambda} f''''(t) .$$

Thus $f(t) = A\cos\sqrt{\lambda}t + B\sin\sqrt{\lambda}t + Ct + D$. (15)

Suppose $f(s)$ is normalized, so that $\int_0^1 f(s)ds = 1$, and let $K = \int_0^1 sf(s)ds$.

Set $\theta = \sqrt{\lambda}$; then $\int_0^1 f(s)ds = 1$ gives

$$\frac{A}{\theta}\sin\theta - \frac{\beta}{\theta}(\cos\theta - 1) + \frac{C}{2} + D = 1 \tag{16}$$

and

$$K = \int_0^1 sf(s)ds = AI_1 + BI_2 + \frac{C}{3} + \frac{D}{2} , \tag{17}$$

where

$$I_1 = \int_0^1 s \cos \theta s \; ds = \frac{\theta \sin \theta + \cos \theta - 1}{\theta^2}$$

$$I_2 = \int_0^1 s \sin \theta s \; ds = \frac{\sin \theta - \theta \cos \theta}{\theta^2} .$$

Substituting $f(t)$ into (13) gives $-Ct - D + 4 - 6t - 6K + 12Kt = 0$ for all t ; thus, equating coefficients, we have

$$- C - 6 + 12K = 0 \quad \text{and} \quad - D + 4 - 6K = 0 .$$

Hence $\frac{C}{3} + \frac{D}{2} = K$, and $C + 2D = 2$.

Thus from (16) we have $A \sin \theta - B(\cos \theta - 1) = 0$, and from (17) we have $AI_1 + BI_2 = 0$. Hence θ must satisfy

$$\frac{\sin \theta}{\cos \theta - 1} = \frac{B}{A} = - \frac{I_1}{I_2} = \frac{1 - \theta \sin \theta - \cos \theta}{\sin \theta - \theta \cos \theta} ; \tag{18}$$

So θ satisfies $2 - \theta \sin \theta - 2 \cos \theta = 0$, by cross-multiplication of (18). Let $\phi = \frac{\theta}{2}$; then $2 - 4\phi \sin \phi \cos \phi - 2[1 - 2 \sin^2\phi] = 0$, and hence

$\sin \phi = 0$ or $\sin \phi - \phi \cos \phi = 0$. Then $\phi_i = \pi i$, $i = 1,2,\ldots$; or alternatively ϕ_k is the solution of $\tan \phi_k = \phi_k$, $k = 1,2,\ldots,$. Finally, $\lambda_i = 4\phi_i^2$, for the first λ-set, and $\lambda_k = 4\phi_k^2$, for the second λ-set.

<u>3.2 Case 2.</u> For Case 2 the test statistic is

$n\Sigma(X_{(i)} - \hat{X}_{(i)})^2/\Sigma(X_{(i)} - \bar{X})^2 = Z_2$. We can take $a = 0$ in the model

$E(X_{(i)}) = a + \beta m_i$, so that this becomes $E(X_{(i)}) = \beta m_i$, with

$\hat{\beta} = \dfrac{\Sigma X_{(i)} m_i}{\Sigma m_i^2}$. Hence $\hat{\beta} - 1 = \dfrac{\Sigma(X_{(i)} - m_i)m_i}{\Sigma m_i^2}$. Similar reasoning to that

for Case 3 gives the asymptotic distribution of Z_2 to be that of

$12\int_0^1 W_2^2(t)dt$ where

$$W_2(t) = Q(t) - 3t\int_0^1 sQ(s)ds. \tag{19}$$

$Q(t)$ is as defined in the previous section, and then $W_2(t)$ is a Gaussian process with mean 0; its covariance function (after some algebra) is

$$\rho_2(s,t) = \min(s,t) - \frac{14}{5}st + \frac{st^3}{2} + \frac{s^3t}{2} . \tag{20}$$

Thus for the weights in the asymptotic distribution of Z_2 , we need eigenvalues of $\lambda\int_0^1 \rho_2(s,t)f(s)ds = f(t)$ Similar steps to those for Case 3 give $f(t) = A\cos\theta t + B\sin\theta t + Ct + D$ with $\theta = \sqrt{\lambda}$, as before. Also, $f(0) \equiv 0$, so $D = -A$, and

$$-f(t) = 3t\int_0^1 sf(s)ds = \frac{f''(t)}{\lambda} . \tag{21}$$

Thus $f''(0) = 0$, so $D = A = 0$. Then, from (21), we have

$$- B\sin\theta t - Ct + 3t[B\int_0^1 s\sin\theta s\, ds + \int_0^1 Cs^2\, ds] \equiv - B\sin\theta t .$$

Hence $\int_0^1 s \sin \theta s \, ds = 0$; thus θ_j is the solution of

$\sin \theta_j - \theta_j \cos \theta_j = 0$, that is, $\tan \theta_j = \theta_j$, $j = 1,2,\ldots$. Finally,

$\lambda_j = \theta_j^2$. These are the weights given in Section 2.

3.3 Asymptotic percentage points. The next step is to calculate the percentage points of, say, $Z_3^* = \Sigma v_i/\lambda_i$ where λ_i are the weights for Case 3. The mean μ_3 of Z_3 is $\int_0^1 \rho_3(s,s)ds = 1/15$. The 80 largest λ_i were found, and Z_3^* was approximated by $S_1 = S^* + T$, where $S^* = \Sigma_1^{80} v_i/\lambda_i$ and $T = \mu_3 - \Sigma_1^{80} \lambda_i^{-1}$. S_1 differs from Z_3^* by $\Sigma_{81}^{\infty} \lambda_i^{-1}(v_i - 1)$ which is a random variable with mean 0 and variance

$$2\Sigma_{81}^{\infty} \lambda_i^{-2} = 2\{\int_0^1\int_0^1 \rho_3^2(s,t)ds\,dt - \Sigma_1^{80} \lambda_1^{-2}\};$$

this value is found to be negligibly small. Thus critical points of Z_j^* are found by finding those of S^* , using Imhof's (1961) method for a finite sum of weighted χ^2 variables, and then adding T .

3.4 Tables. Tables 1 and 2 give percentage points for Z_2 and Z_3 respectively. Those for n finite have been obtained by Monte Carlo sampling. The last line in each table contains the asymptotic points. Table 1 also gives points for a modification of Z_2 , called Z_{2A} . This is the statistic (using the terminology for Case 2 in Section 2) $Z_{2A} = n\Sigma\{X_{(i)} - \hat{X}_{(i)}\}^2/\Sigma X_{(i)}^2$. This uses the quantity $\Sigma X_{(i)}^2/n$ to eliminate the square of the scale instead of the sample variance. This is a natural denominator in Case 2 with a = 0, where the model is $E(X_{(i)}) = \beta m_i$. (If a is not zero, the new variable $X'_{(i)} = X_{(i)} - a$ should be used instead of $X_{(i)}$. The asymptotic points for Z_{2A} are 0.25 times those for Z_2. An advantage in using Z_{2A} is that the statistic is much less variable for small n. For Z_3, Table 2 has already been produced in Stephens (1986), although with less accurate points; there will be negligible difference in practical use.

3.5 Use of the Tables with censored data. Suppose origin and scale are both unknown (Case 3), and the data is censored at both ends. Thus $n^* = r - k + 1$ observations are available, consisting of all those between $X_{(k)}$ and $X_{(r)}$. $R(X,T)$ may be calculated, using the usual formula, but with sums for i from k to r, and with $T_i = i/(n+1)$ or $T_i = i$, or even $T_1, T_2, \ldots$ equal to $1, 2, \ldots, n^*$, these latter values for T_i being possibilities because $R(X,m)$ is scale and location invariant. Then $n^*\{1 - R^2(X,T)\} = Z(X,T)$ will be referred to Table 3, using the values for sample size n^*.

3.6 Example. It is well-known that if times Q_i; $i = 1,2,\ldots,n$ represent times of random events, occurring in order with the same rate, the $Q_{(i)}$ should be proportional to uniform order statistics $U_{(i)}$. Thus the $Q_{(i)}$ may be regressed against $i/(n+1)$ or equivalently against i as described above, to test that the events are random. Suppose $Q_{(9)}, Q_{(10)}, \ldots, Q_{(20)}$ represent a subset of such times, denoting times of breakdown of an industrial process. We wish to test that these are uniform; times $Q_{(1)}$ to $Q_{(8)}$ have been omitted because the process took time to stabilize and these are not expected to have occurred at the same rate as the later times. The times $Q_{(9)}, \ldots, Q_{(20)}$ are 82, 93, 120, 135, 137, 142, 162, 163, 210, 228, 233, 261. The value of $Z(Q,T) = 12\{1 - R^2(Q,T)\}$

$= 0.464$. Reference to Table 2 at line $n = 12$ show that there is not significant evidence, at the 10% level, to reject the hypothesis of uniformity.

4. THE CORRELATION COEFFICIENT: GENERAL CASE.

4.1 The general case. We now discuss, in a non-rigorous fashion, the distribution of $Z(X,m)$ for the general test of H_0 given in (5). $F_o(x)$ is assumed to be a continuous distribution, and the sample can be left- and right-censored. Thus we observe $X_{(k)} < \dots < X_{(r)}$ from a sample of size n from the distribution $F_o(x)$. We can assume the sample comes from $F_o(x)$ with $\alpha = 0$ and $\beta = 1$, that is, from $F(\cdot)$ although (3) is fitted without this knowledge. Suppose $f(x)$ is the density corresponding to $F(x)$. Then using $H_i = F^{-1}(\frac{i}{n+1})$ we have

$$Z(X,m) = n\{1 - R^2(X,m)\} = \frac{\Sigma_k^r\{X_i - H_i - \hat{\alpha} - (\hat{\beta} - 1)\}^2}{\frac{1}{n}\Sigma_k^r(H_i - \bar{H})^2}.$$

Define $p = (k-1)/n$ and $q = r/n$ and let $q^* = F^{-1}(q)$ and $p^* = F^{-1}(p)$. Also, let

$$Y(t) = Q(t) - \int_p^q Q(s)ds - \frac{\{F^{-1}(t) - \mu\}}{\sigma}\int_p^q \{\frac{F^{-1}(s) - \mu}{\sigma}\}Q(s)ds,$$

where

$Q(t) = \sqrt{n}\,\{X_{([nt])} - F^{-1}(t)\}$, and parameters μ and σ are given by

$$\mu = \int_p^q F^{-1}(s)ds = \int_{p^*}^{q^*} xf(x)dx ,$$

and

$$\sigma^2 = \int_p^q (F^{-1}(s))^2 ds - \mu^2 = \int_{p^*}^{q^*} x^2 f(s) dx - \mu^2 .$$

The process $Q(t)$ is close to a Gaussian process with mean 0 and covariance

$$\rho_o(s,t) = \frac{\min(s,t)}{f(F^{-1}(s))f(F^{-1}(t))} .$$

The process $Y(t)$ is then close to a Gaussian process with mean 0 and covariance

$$\rho(s,t) = \rho_o(s,t) - \psi(s)\int_p^q \psi(u)\rho_o(u,t)du - \psi(t)\int_p^q \psi(u)\rho_o(s,u)du$$

$$- \int_p^q \rho_o(u,t)du - \int_p^q \rho_o(s,u)du + \int_p^q\int_p^q \rho_o(u,v)dudv$$

$$+ \psi(s)\psi(t) \int_p^q\int_p^q \rho_o(u,v)\psi(u)\psi(v)dudv + (\psi(s) + \psi(t))\int_p^q\int_p^q \rho_o(u,v)\psi(u)dudv$$

where $\psi(s) = \frac{F^{-1}(s) - \mu}{\sigma}$.

The denominator of $Z = n\{1 - R^2\}$, where we write Z for $Z(X,m)$ and R^2 for $R^2(X,m)$, i. then close to σ^2, and the numerator is close to $T = \int_p^q Y^2(t)dt$. Thus the asymptotic theory now depends on the behaviour of T . It appears generally that this behaviour is determined by that of $\int_p^q Q^2(t)dt$. There are 3 cases in practice, which we label Cases A, B and C. Define

$$J_1 = \int_p^q \int_p^q \rho_0^2(s,t)\,ds\,dt$$

and

$$J_2 = \int_p^q \rho_o(t,t)\,dt \; .$$

Case A. In this case suppose $J_1 < \infty$ and $J_2 < \infty$. Then we have

$$Z = n(1 - R^2) \Rightarrow \frac{1}{\sigma^2} \Sigma_1^{\infty} \; v_i/\lambda_i$$

where v_i are independent χ_1^2 variables and λ_i are eigenvalues of

$f(s) = \lambda \int_p^q \rho(s,t) f(t)\,dt$. (The sum $\Sigma\lambda_i^{-1}$ will be $< \infty$).

Case B. Suppose $J_1 < \infty$ but $J_2 = \infty$. Then there exists $a_n \to \infty$ such that $Z - a_n \equiv n(1 - R^2) - a_n \Rightarrow \frac{1}{\sigma^2} \Sigma_1^{\infty} \lambda_i^{-1}(v_i - 1)$, where the λ_i and v_i are as defined above. (In this case $\Sigma\lambda_i^{-1} = \infty$.)

Case C. For this case suppose both integrals J_1 and J_2 are infinite. Then there exist constants a_n, b_n, such that

$$\frac{Z - a_n}{b_n} \equiv \frac{n(1 - R^2) - a_n}{b_n} \Rightarrow N(0,1) \, .$$

4.3 Examples.

1. The exponential distribution.

For $q = 1$ we have case C; $a_n = \log n$, and $b_n = (\log n)^{\frac{1}{2}}$, so that

$$\frac{n(1 - R^2) - \log n}{2\sqrt{(\log n)}} \Rightarrow N(0,1).$$

For $q < 1$ we are in Case A and the distribution is a sum of weighted chi-squared variables.

2. The uniform test (discussed above).

For any p or q Case A applies and $(r - k+1)(1 - R^2)$ has the same limiting distribution *regardless of* p, q .

3. The normal test.

For $p = 0$ or $q = 1$ or both we get Case B .

For $p > 0$, $q < 1$ we get Case A.

4. The Logistic test: $F(w) = 1/(1 + e^{-w})$, $-\infty < \omega < \infty$.

For $p = 0$ or $q = 1$ or both we get Case C.

For $p > 0$ and $q < 1$, we get Case A. The logistic test is thus similar to the exponential test.

5. Test for the Extreme Value distribution I: $F(w) = 1 - e^{-e^{w}}$, $-\infty < \omega < \infty$.

For $p = 0$, we get Case C.

For $p > 0$ we get Case A.

6. Test for the Extreme Value distribution II: $F(w) = e^{-e^{-w}}$, $-\infty < \omega < \infty$.

For $q = 1$, we get Case C.

For $q < 1$, we get Case A.

4.4 Discussion. The discussion above is somewhat imprecise. When p is 0 or q is 1 there are technical details which have been glossed over. For the distributions we have studied however the criteria given in Cases A, B and C lead to the correct answer for asymptotic distributions of $Z(X,m) = n\{1 - R^2(X,m)\}$.

Table 1. Critical Points for Z_2 and Z_{2A}.

Upper tail significance level (percent)

	n	50	25	15	10	5	2.5	1
Z_2	4	0.690	1.240	1.94	3.47	8.67	20.3	47.0
	6	0.763	1.323	1.89	2.59	4.74	8.49	17.0
	8	0.806	1.364	1.85	2.37	3.78	6.29	11.4
	10	0.832	1.388	1.88	2.34	3.40	5.30	8.9
	12	0.848	1.407	1.89	2.33	3.27	4.80	7.8
	18	0.877	1.438	1.91	2.32	3.12	4.26	6.3
	20	0.881	1.444	1.92	2.32	3.10	4.18	6.0
	40	0.907	1.470	1.93	2.32	3.03	3.82	5.1
	60	0.916	1.480	1.93	2.32	3.00	3.73	4.9
	80	0.920	1.485	1.94	2.32	2.99	3.71	4.9
	100	0.922	1.488	1.94	2.32	2.98	3.70	4.8
	∞	0.932	1.497	1.94	2.31	2.98	3.67	4.6
Z_{2A}	4	0.140	0.245	0.333	0.411	0.545	0.707	1.010
	6	0.166	0.287	0.379	0.467	0.616	0.796	1.065
	8	0.184	0.307	0.403	0.494	0.648	0.830	1.089
	10	0.193	0.320	0.420	0.512	0.670	0.848	1.102
	12	0.200	0.330	0.432	0.523	0.683	0.861	1.111
	18	0.209	0.346	0.452	0.543	0.705	0.882	1.121
	20	0.212	0.349	0.455	0.547	0.708	0.886	1.124
	40	0.224	0.362	0.472	0.563	0.727	0.903	1.138
	60	0.228	0.367	0.477	0.568	0.734	0.909	1.146
	80	0.229	0.369	0.479	0.570	0.736	0.911	1.149
	100	0.230	0.370	0.480	0.572	0.737	0.912	1.150
	∞	0.233	0.374	0.485	0.578	0.744	0.917	1.155

Table 2. Critical Points for Z_3 .

n	0.5	0.25	0.15	0.10	0.05	0.025	0.01
4	0.344	0.559	0.734	0.888	1.089	1.238	1.388
6	0.441	0.703	0.901	1.053	1.325	1.590	1.918
8	0.495	0.792	1.000	1.163	1.474	1.739	2.100
10	0.535	0.833	1.068	1.245	1.532	1.846	2.294
12	0.560	0.864	1.093	1.280	1.608	1.918	2.360
18	0.605	0.940	1.147	1.348	1.672	2.008	2.503
20	0.610	0.960	1.200	1.370	1.680	2.025	2.520
40	0.640	0.980	1.215	1.396	1.732	2.076	2.580
60	0.648	0.988	1.227	1.410	1.750	2.092	2.590
80	0.658	0.997	1.228	1.418	1.760	2.104	2.610
∞	0.666	0.992	1.234	1.430	1.774	2.129	2.612

References

1. Durbin, J., (1973). Distribution theory for tests based on the sample distribution function. Regional Conference Series in Appl. Math., 9. Philadelphia: SIAM.
2. Gerlach, B., (1979). A consistent correlation-type goodness-of-fit test; with application to the two-parameter Weibull distribution. Math. Operations-forsch. Statist. Ser. Statist. 10, 427-452.
3. Imhof, J.P., (1961). Computing the distribution of quadratic forms in normal variables. Biometrika, 48, 419-426.
4. Sarkadi, K., (1975). The consistency of the Shapiro-Francia test. Biometrika 62, 445-450.
5. Stephens, M.A., (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann. Statist. 4, 357-369.
6. Stephens, M.A., (1986). Tests based on regression and correlation. Chapter 5 of Goodness-of-fit Techniques , (R.B. D'Agostino and M.A. Stephens, eds.) New York: Marcel Dekker.
7. Watson, G.S., (1961). Goodness of fit tests on a circle. Biometrika 48, 109-14.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 436	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Tests Of Fit Based On The Correlation Coefficient		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Richard A. Lockhart and M. A. Stephens		8. CONTRACT OR GRANT NUMBER(s) N00014-89-J-1627
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE October 4, 1990
		13. NUMBER OF PAGES 27
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Goodness-of-fit, Correlation coefficient, Correlation tests.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE FOLLOWING PAGE.

TECHNICAL REPORT NO. 436

20. ABSTRACT

In testing fit of a sample to a specific family of distributions F, probability-plots use a plot of the order statistics of the sample against a set of suitable constants. These are chosen so that if the parent population is one of the family F, the resulting plot should appear like a straight line. Members of F may contain unknown location and/or scale parameters. Historically, the judgment of a straight line fit was usually made by eye. In this article we propose the correlation coefficient R between the sample values and the constants as a test statistic. Asymptotic properties of R are derived in general, and tables produced to make the test for the case where the proposed distribution is uniform.